Mathematics 376 – Probability and Statistics II
Lab Project 4 – A Regression Case-Study
April 30 - May 3, 2004

*Background*

In class, we have discussed the basic methods for obtaining the least-squares estimators for the coefficients $\beta_i$ in simple linear models

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

and techniques for hypothesis testing concerning the $\beta_i$ (see page 553 in the text for a convenient summary). In this lab project, we will work through a "case-study" of how these methods might be used in a realistic statistical study.

*The "Story"*

Heat treating is often used to increase the carbon content of the metal in machine parts such as gears. The thickness of the carbonized layer is considered an important contributor to the overall reliability of the part. A lab test is performed on some of the gears in each batch that is manufactured, in which the gear is sliced apart, then soaked in a chemical bath for some length of time. The carbon layer thickness is measured at the end of the soaking period. The following table gives the data collected in one such test:

| Soak time($x$) | C thickness($y$) | Soak time($x$) | C thickness($y$) |
|---|---|---|---|
| 0.58 | .013 | 1.17 | .021 |
| 0.66 | .016 | 1.17 | .019 |
| 0.66 | .015 | 1.17 | .021 |
| 0.66 | .016 | 1.20 | .025 |
| 0.66 | .015 | 2.00 | .025 |
| 0.66 | .016 | 2.00 | .026 |
| 1.00 | .014 | 2.20 | .024 |
| 1.17 | .021 | 2.20 | .025 |
| 1.17 | .018 | 2.20 | .024 |
| 1.17 | .019 | | |

From this data, the foundry managers are concerned that the accuracy of the carbon thickness measurements might be affected by the length of the chemical soak, since there seems to be a slight(?) upward trend to the thickness measurements as the soak time is increased. Imagine your lab group is an internal statistical research group employed by the foundry that makes these gears. Your job is to determine whether this trend is actually meaningful (statistically significant), or whether it can be "explained away" by citing randomness in the manufacturing and testing procedures. You will also see how to study the form of the actual functional relationship between the soak time and the carbon thickness measurement.

A) Start your analysis on an intuitive level – compute the equation of the regression line that best fits the data, and plot the data points together with that line.

B) If the variations here were really independent of $x$ = the soak time before the measurement, then we would expect $\beta_1 = 0$ in the model (1) – the regression line would be approximately horizontal. Your estimated value of $\beta_1$ in part A should be quite small, so it may not be obvious whether $\beta_1 = 0$ is consistent with the data. But of course we have a lot of information for making inferences about the regression coefficients. So test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$ using an appropriate test. Give the $p$-value of your test. Also give a 99% confidence interval for $\beta_1$.

C) Report the conclusion from the test you carried out in part B in complete English sentences in a short paragraph that *even the managers of the foundry – your bosses – could understand.* (In this paragraph *don't use jargon* like "reject the null hypothesis at the $\alpha = .05$ level." Remember, the bosses don't know anywhere near as much statistics as you do, and just want to be told clearly what your results are and what they mean for the foundry!)

D) Now the bosses probably don't care much about this next part, but you have become intrigued by this data and you want to figure out whether the linear model (1) is actually even a good one to use for this problem. Carry out an *analysis of variance test* for linearity of regression using this data. The null hypothesis now is that a linear model is a good fit for this data; the alternative hypothesis is that it is not. This will involve an $F$-test using the computed lack of fit mean square as we discussed in class. State your conclusion, citing the $p$-value of your test.

E) Might a different functional form $Y = f(x) + \epsilon$ give a better fit for this data? This is a huge question, and leads into a very interesting part of statistics, called *multiple regression*, that we unfortunately do not have the time to explore since we are at the end of the course. Here's just a simple taste of some of the ideas. For example, we can ask whether a quadratic polynomial function of $x$ might give a better model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{2}$$

The computation of the least squares estimators for $\beta_0$, $\beta_1$, and $\beta_2$ can be derived much as we did for the the simple model (1). But an even better way uses *matrices* and some ideas from linear algebra. For the purposes of the following Maple commands, we will assume that the data from the table above has been entered in two lists $XL$, and $YL$. Enter the following commands:

```
with(linalg):
Y:=transpose(matrix([[seq(YL[i],i=1..17)]])):
X:=transpose(matrix([[seq(1,i=1..17)],[seq(XL[i],i=1..17)],
    [seq(XL[i]^2,i=1..17)]])):
```

which create two matrices – one $17 \times 1$ column matrix containing the entries of $YL$ – the thickness measurements, and a $17 \times 3$ matrix whose first column is all 1's, second is the $x_i$, and third is the $x_i^2$. The idea is that the normal equations for computing the least squares estimators of $\beta_i$ can be written as the system of linear equations:

$$X^t X b = X^t Y$$

To find the least squares estimators, we solve this system of linear equations for the vector $b$. In Maple, we can do that as follows:

```
XtX:=multiply(transpose(XM),XM);
XtY:=multiply(transpose(XM),YM);
b:=linsolve(XtX,XtY);
```

Here $b$ is the 1-column matrix containing the least squares estimators of $\beta_i$ in (2): $\widehat{\beta}_0$ is `b[1,1]`, $\widehat{\beta}_1$ is `b[2,1]`, and $\widehat{\beta}_2$ is `b[3,1]`. Carry out these computations, and plot the regression parabola

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$$

together with the data points and the regression line you computed before.

F) On page 587 of the text, you will find a discussion of tests of hypotheses in multiple regression. Adapt that to figure out how to test the null hypothesis $\beta_2 = 0$ versus the alternative that $\beta_2 \neq 0$. What is the conclusion of this test? Is the model (2) really better than (1) here??

*Assignment*

One Maple worksheet from each lab group. Due by the end of the day on Wednesday, May 5 in Swords 335.